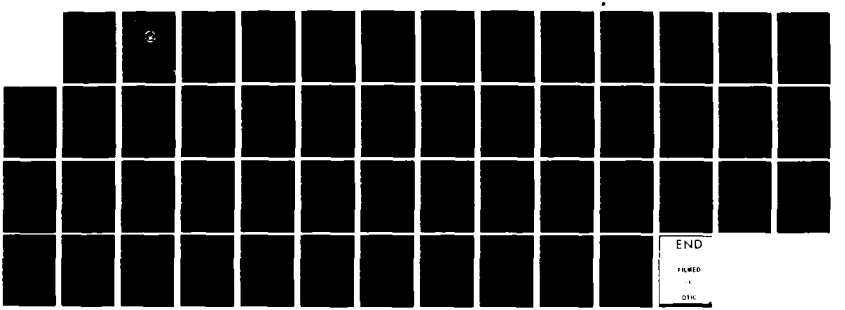
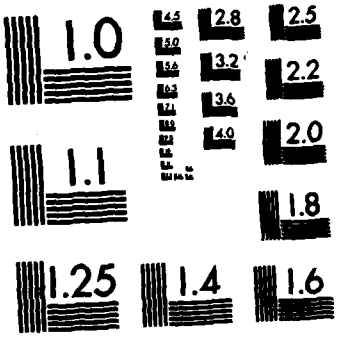
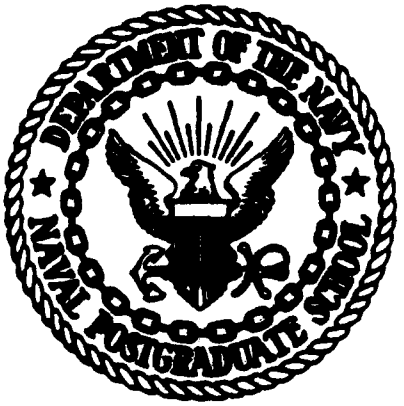
AD-A124 586 AN OPTIMIZATION APPROACH TO STUDY THE DISTRIBUTION OF THE COEFFICIENT OF DETERMINATION R2(U) NAVAL POSTGRADUATE SCHOOL MONTEREY CA J D VICK OCT 82 1/1

UNCLASSIFIED F/G 12/1 NL





NAVAL POSTGRADUATE SCHOOL
Monterey, California



THESIS

AN OPTIMIZATION APPROACH TO STUDY THE DISTRIBUTION OF THE COEFFICIENT OF DETERMINATION, R^2

by

Jeffery Donald Vick

October 1982

Thesis Advisor: G. T. Howard

Approved for public release; distribution unlimited.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO. AD-A124586	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) An Optimization Approach to Study the Distribution of the Coefficient of Determination, R^2		5. TYPE OF REPORT & PERIOD COVERED Master's thesis; October 1982
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Jeffery Donald Vick		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Postgraduate School Monterey, California 93940		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Naval Postgraduate School Monterey, California 93940		12. REPORT DATE October 1982
		13. NUMBER OF PAGES 50
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

"Coefficient of Determination" — R^2
regression — optimization
costing — prediction
pricing

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The empirical model builder utilizing regression techniques frequently relies on the coefficient of determination, R^2, to measure 'goodness of fit'. Costing and pricing analysts using such variable selection techniques frequently encounter inflated R^2 values. This paper examines the space within which the regression model operates and presents practical optimization algorithms to help assess the amount of confidence that can be

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

#20 - ABSTRACT - (CONTINUED)

placed in R^2 for a particular set of candidate predictor variables. The algorithms describe a technique using linear programming to find the lowest value of R^2 possible using the given set of data.

Approved for public release; distribution unlimited.

An Optimization Approach to Study the Distribution of the Coefficient of Determination, R^2

by

Jeffery Donald Vick
Captain, United States Marine Corps
B.S., University of Minnesota, 1976

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL

October 1982

Author: Jeffery D. Vick

Approved by: G. T. Howard
Thesis Advisor

Michael D. Foregger
Second Reader

Kneale T. Marshall
Chairman, Department of Operations Research

W. M. Woods
Dean of Information and Policy Sciences

ABSTRACT

The empirical model builder utilizing regression techniques frequently relies on the coefficient of determination, R^2, to measure 'goodness of fit'. Costing and pricing analysts using such variable selection techniques frequently encounter inflated R^2 values. This paper examines the space within which the regression model operates and presents practical optimization algorithms to help assess the amount of confidence that can be placed in R^2 for a particular set of candidate predictor variables. The algorithms describe a technique using linear programming to find the lowest value of R^2 possible using the given set of data.

TABLE OF CONTENTS

LIST OF FIGURES

I. INTRODUCTION

A. PROBLEM MOTIVATION

Parametric cost estimation is a management tool used to aid in the prediction of the cost of a proposed system. It involves predicting the cost (dependent variable) of a system by means of explanatory (independent) variables such as system characteristics or performance requirements. This procedure is based on the premise that the cost of a system is related in a quantifiable way to the system's physical and performance characteristics. The expression of this quantifiable relationship is in the form of an estimating equation derived through statistical regression analysis of historical cost data on systems which are, more-or-less, analogous to the proposed system. Since parametric cost estimates can be developed during the concept formulation stage of the acquisition process before engineering plans are finalized, these estimates can be used by management to:

(1) Identify possible cost/performance tradeoffs in the design effort.

(2) Provide a basis for cost/effectiveness review of performance specifications.

(3) Provide information useful in the ranking of competing alternatives.

(4) Suggest a need for investigating new alternatives.

Cost overruns have been prevalent in the acquisition process for new weapon systems making cost estimation a very

important problem for all components of the Department of Defense.

To combat this problem, the Department of Defense has issued directives to employ independent parametric cost estimation. Publications such as Reference [1] have appeared which give step by step methodology for the development of a parametric cost estimate.

Regression problems faced by costing and pricing analysts in these situations are inherently difficult for two fundamental reasons [Ref. 2]:

(1) The number of observations is usually small compared with the number of system characteristics which are candidate components of the regression equation.

(2) The available data is not produced by employing an efficient experimental design.

Under these circumstances, it has been shown that the use of variable selection techniques may result in regression equations which yield inflated R^2 values whose statistical significance cannot be tested using the F-test.[1]

In general parametric cost estimation, an analyst should not blindly trust the regression equation resulting from his analysis. To measure the 'goodness of fit', the analyst can use such statistics as R^2 and F. (As noted earlier, however, regression models for cost estimation often do not allow use of F.) There are few hard and fast rules for assessing the

---

[1]This is the case when R^2 is not significant, and some, but not all, of the β_i are significant.

usefulness of such a model. This is especially true of models that result from the application of a variable selection technique in order to obtain a 'best' prediction equation. The R^2 statistic in these situations may not give a meaningful indication of the model's applicability.

The purpose of this paper is to investigate the coefficient of determination, R^2, used in best subset regression analysis. The solution algorithms presented in this paper provide a practical method to help assess the confidence placed by the empirical model builder in R^2 for a regression upon a particular set of exogenous data. It may contribute to the theoretical foundation for the understanding of regression models, whose properties are not fully understood.

B. *STATEMENT OF THE PROBLEM*

Suppose the analyst selects n-independent observations on p-predictor (candidate) variables and one dependent variable. The goal of the analysis is to determine the k-variable regression equation which maximizes the coefficient of determination for various values of k. The difficulty with this analysis is assessing the statistical significance of R^2 for a given value of k.

The n-independent observations, mathematically, span a n-dimensional finite vector space (call it E^n). The regression procedure projects the dependent variable, Y, onto subspaces within E^n looking for the best fit (prediction). How the subspaces are oriented in E^n therefore dictates the quality

of fit that can be obtained. The subspaces for E^n are determined by the candidate predictor variables; each predictor variable's observed values representing a vector in E^n, and each combinatorial set of the p-predictor variables spanning a subspace. Obviously, there are $\binom{p}{k}$ possible k-variable prediction equations. Therefore, the analyst's selection of the p-variables to be used as candidates determines which subspaces are available for the regression procedure to consider in best subset selection.[2]

Wallenius [Ref. 3] in trying to gather information on the unknown distribution of R^2 asks, "How well do the $\binom{p}{k}$ subspaces spanned by all the subsets of columns of X 'fill' E^n?" (Where X is the (n x p) matrix of n-observations on the system's p-characteristics.) In other words, using the candidate predictor variables selected by the analyst, will the highest R^2 value obtainable through regression differ very much from the worst possible?

Wallenius characterized this problem mathematically by defining the 'coefficient of fill' (COF) as follows:

$$R^2_{min}(X,k) = \min_{\underline{Y}} \max_{\{i_1,\ldots i_k\}} R^2_{\underline{Y};(x_{i_1},x_{i_2},\ldots x_{i_k})} .$$

This formula can be interpreted to ask, "If given some set of candidate predictor variables, what is the worst $\underline{Y}$ that

---

[2]The total number of exogenous variables that Y is regressed upon is p.

can be predicted?" Where 'worst' can be identified by the lowest R^2 value obtainable. Thus, a lower bound for R^2 is also obtained by answering this.

The problem is extremely difficult to solve directly. This paper proposes an algorithm for solving this problem using optimization with a surrogate objective function. The number of optimizations required to obtain a global solution is exponentially related to p (the number of predictor variables). Thus, in the area of p = 14, enumeration begins to become economically infeasible as a solution technique, and hence, an algorithm is proposed to search the area about some minimum point. This latter algorithm cannot guarantee a global solution, so it must be considered local in nature. Whether such a local solution is useful requires further research; usefulness may be directly dependent upon the empirical nature of the data.

II. SUPPORTING CONCEPTS

A. GENERAL LINEAR REGRESSION USING LEAST SQUARES

A general and frequently used linear model is the 'multiple linear regression model'. It can be represented as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_n x_{in} + \varepsilon_i .$$

The variable represented by y is the variable of interest (i.e., to be predicted). The variables represented by the x_j are associated with y and may influence the behavior of y. Thus, mathematically y is called the dependent variable (endogenous) and the x variables are called independent variables (exogenous). Statistically, this model is referred to as the regression of y on the x variables. The coefficients β are referred to as 'partial regression coefficients' and they specify the linear functional relationship between the independent variables and the dependent variable. Mathematically, the β_j are the partial derivatives of the functional relationship $\partial y/\partial x_j$. Thus, a β_j indicates the change in the dependent variable y corresponding to a unit change in the independent variable x_j (all other independent variables held fixed).

There are various criteria used in regression, however the formulation of interest for this paper is based upon the least

squares criteria. Use of least squares to derive the formula for estimating Y is shown below.

Using matrix notation, the regression model can be written as:

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon} ,$$

where

$$\underline{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} , \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & & & & \\ 1 & & & & \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} ,$$

$$\underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{m+1} \end{bmatrix} , \quad \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

(a) X is the n x (m+1) matrix of n-observations on m-independent (x) variables plus a dummy variable.

(b) $\underline{Y}$ is the column vector of the n-observed values of Y.

(c) $\underline{\beta}$ is the column vector of the m+1 partial regression coefficients (whereas, $\hat{\underline{\beta}}$ is the vector of estimated regression coefficients).

Formulating the least squares

$$\min \underline{\varepsilon}^T\underline{\varepsilon} = (\underline{Y}-X\underline{\beta})^T(\underline{Y}-X\underline{\beta}) ,$$

taking the derivative with respect to $\underline{\beta}$

$$\frac{\partial}{\partial\beta}[\underline{Y}^T\underline{Y} - 2\underline{Y}^TX\underline{\beta} + \underline{\beta}^TX^TX\underline{\beta}] = 0 ,$$

and then solving for $\underline{\beta}$, we get the estimate

$$\hat{\underline{\beta}} = (X^TX)^{-1}X^T\underline{Y} .$$

B. STANDARDIZATION AND NORMALIZATION OF VECTORS

The regression coefficients of the linear model are functions of the units of measurement of the variables. The magnitudes of coefficients are influenced by choices of units of measurement. Thus, a tantamount scaling problem to that experienced in linear programming exists. This scaling problem is avoided by use of 'standardized regression coefficients'.

Standardized regression coefficients are the end result when the variables which they are estimated from have been transformed to unit variance.

Consistent with previous notation, and for use further on in this paper, an alternative form to obtain 'standardized' predictions of the dependent variable y_i for all i (represented as $\underline{Y}^*$) is:

$$\underline{Y}^* = \frac{1}{\sigma_Y}(\underline{Y} - \bar{y}\,\mathbf{1}), \quad \text{or} \quad \underline{Y} = \frac{1}{\sigma_Y}(\underline{Y} - \bar{\underline{Y}})$$

(a) (*) denotes 'standardized'.

(b) $\mathbf{1}$ denotes a (n x 1) column vector of ones.

Of great importance is that a linear transformation has been performed, and that the intercept of the regression equation is zero (i.e., the equation passes through the origin). Each partial regression coefficient indicates how many standard deviation changes in y are associated with one standard deviation change in the corresponding x (all other x_j held fixed). Also, a mathematical characteristic is that

$$\sum_{i=1}^{n} y_i = 0$$

or, equivalently,

$$\mathbf{1}^T\underline{Y}^* = 0 .$$

Mathematically, normalization is a linear transformation that takes any given vector and converts its length to unit length (length in multidimensional vector spaces discussed in the next section). A vector of unit length is said to have a norm = 1. Any arbitrary vector can be transformed into a unit vector by dividing it by its norm.

A vector with unit length can be depicted as follows:

$$\sum_{i=1}^{n} y_i = 1 , \quad \text{or} \quad \underline{Y}^T\underline{Y} = 1 .$$

C. LENGTH, ANGLE, AND COSINE FUNCTION IN MULTIDIMENSIONAL VECTOR SPACES

Using the concept of inner products, length or magnitude of a vector can be defined. In this context, the length is referred to as the 'norm' of the vector.

The norm of the arbitrary vector $(x_1,x_2,\ldots,x_n)$ in R^n is denoted by

$$||(x_1,x_2,\ldots,x_n)|| = \sqrt{(x_1,x_2,\ldots,x_n)(x_1,x_2,\ldots,x_n)} .$$

In matrix notation,

$$||\underline{X}|| = (\underline{X}^T\underline{X})^{1/2} .$$

Thus, a normalized vector would be created as follows:

$$\hat{\underline{X}} = \underline{X} \cdot \frac{1}{||\underline{X}||} = \frac{\underline{X}}{(\underline{X}^T\underline{X})^{1/2}} .$$

To obtain the cosine of the angle between two vectors, one normalizes each vector and then takes their inner product. If u and v are vectors in R^n, the cosine of the angle can be defined as:

$$\cos\theta = \frac{1}{||u||} u \cdot \frac{1}{||v||} v$$

Two vectors are orthogonal if and only if their inner product is zero. This means the angle between them is 90°, $\cos\theta = 0$, and $v \cdot u = 0$.

D. PROJECTIONS

Let u and v be vectors with angle α between them. The scalar projection (or component) of v in the direction of u is defined to be $||v||\cos\alpha$. Geometrically, it can be visualized as in figure one.[3]

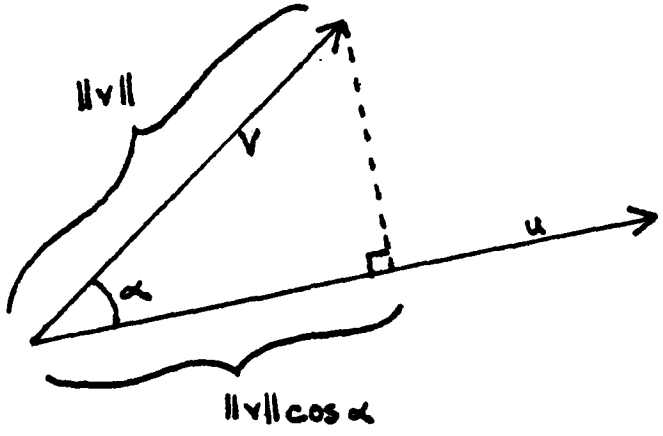


Figure 1. Scalar Projection of V onto U

Alternatively, computation is easier if written as:

$$\text{scalar projection of v onto u} = v \cdot \frac{u}{||u||} .$$

This is the inner product (dot product in E^n) of v with the unit vector in the direction of u.

---

[3]Note that the dashed line is perpendicular to u.

The vector projection is the scalar projection times the unit vector in the direction of u. So, the vector projection of v onto u is written:

$$\text{proj}_u v = ||v||\cos\alpha \cdot \frac{u}{||u||} = (v \cdot \frac{u}{||u||})\frac{u}{||u||} \cdot \quad \cdot$$

III. PROBLEM FORMULATION ONE

Recalling from Chapter II the expression for estimating the regression coefficients,

$$\underline{\hat{\beta}} = (X^T X)^{-1} X^T \underline{Y} ,$$

there is a unique solution for $\underline{\hat{\beta}}$ if the matrix $X^T X$ is non-singular, that is, if it is full rank. This is the case if X has n-independent columns, since the columns of $X^T X$ and the rows of X span the same space [Ref. 4].

Thus, given $\simeq$ (n x p) matrix X of full rank and $n < p$, a finite dimensional vector space E^n can be defined where n-linearly *independent column vectors* span the space (i.e., form a basis). Extending linear algebra concepts to the linear regression model of Section II.A, a geometric interpretation is that k columns (where $k \leq n$) of the X matrix span a k-dimensional subspace in the n-dimensional space E^n. Further, the least squares procedure, through 'best' subset selection, will select amongst the $\binom{p}{k}$ columns of X a k-dimensional subspace of E^n to predict the vector of dependent variables ($\underline{Y}$) such that $\cos^2\theta = R^2$ is maximized (θ is minimized); where θ is the angle between $\underline{Y}$ and its orthogonal projection onto a candidate subspace.

A. RESTATEMENT OF 'COEFFICIENT OF FILL' (COF)

Consider the matrix X. Assume it has rank = n. Next, require that $\underline{Y}$, the vector of dependent variables be obtained through standardized regression coefficients (see Section II.B); that is,

$$\underline{Y}^* = \frac{1}{\sigma_Y}(\underline{Y} - \overline{\underline{Y}}) \ .$$

This requirement causes no loss in generality since,

$$R^2_{\underline{Y}^*;(x_1,\ldots x_k)} = R^2_{\underline{Y};(x_1,\ldots x_k)} \ .$$

Further, by requiring $\underline{Y}^*$ to be a unit vector (normalized), a unique $\underline{Y}$ is specified (i.e., all other vectors in the direction of $\underline{Y}$ will be a scalar multiple of the normalized vector); that is

$$\underline{Y}^{*T}\underline{Y}^* = 1 \ .$$

The 'coefficient of fill' (defined in Section I.B) now becomes [Ref. 3]:

$$R^2_{min}(X,k) = \min_{\substack{1^T\underline{Y}^*=0 \\ \underline{Y}^{*T}\underline{Y}^*=1}} \max_{\{i_1,\ldots i_k\}} R^2_{\underline{Y}^*;(x_{i_1},\ldots x_{i_k})} \ .$$

B. TRANSFORMATION OF THE COF TO A QUADRATIC FORM

Wallenius [Ref. 3] showed that the R^2 portion of the COF can be transformed into a quadratic form. Such a transformation can be done as follows.

Identify arbitrarily any combination of the $\binom{p}{k}$ columns of X. Let the i^{th} such combination be called D_i. Recall that the columns of D_i define a subspace in E^n and any vector in E^n can be projected onto that subspace. Assume that D_i has rank k.

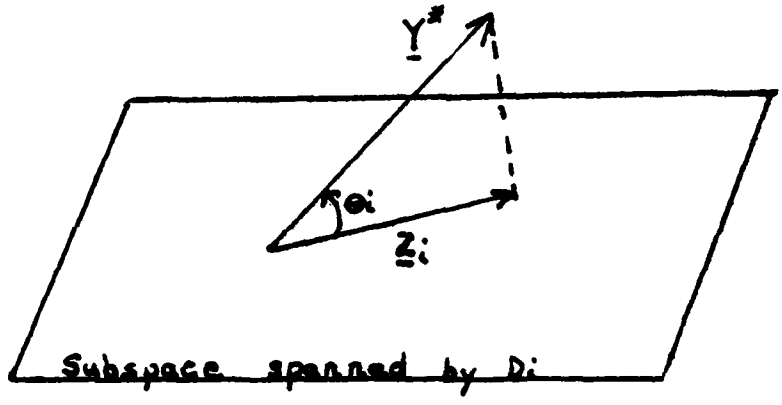


Figure 2. Subspace Spanned by D

From linear algebra (vector projections) and figure two, it follows that

$$\underline{Z}_i = \text{proj}_{D_i} \underline{Y}^* = \sum_{i=1}^{k} (\underline{Y}^* \cdot \hat{\underline{X}}_i) \hat{\underline{X}}_i$$

where $\hat{\underline{X}}_1, \ldots, \hat{\underline{X}}_k$ is an orthonormal basis for D_i.

From least squares regression, an alternate calculation for $\underline{Z}_i$ is as follows:

$$\underline{Z}_i = D_i(D_i^T D_i)^{-1} D_i^T \underline{Y}^* .$$

Recalling that

$$\hat{\underline{\beta}}_i = (D_i^T D_i)^{-1} D_i^T \underline{Y}^* ,$$

and that the 'generalized inverse' of a matrix is

$$D_i^- = (D_i^T D_i)^{-1} D_i^T ,$$

the calculation of $\underline{Z}_i$ can be stated as:

$$\underline{Z}_i = D_i \hat{\underline{\beta}}_i \quad \text{where} \quad \hat{\underline{\beta}}_i = D_i^- \underline{Y}^* .$$

In Section II.C, the cosine of the angle between two vectors was defined. To calculate the angle between $\underline{Y}^*$ and its orthogonal projection $\underline{Z}_i$, use

$$\cos \theta_i = \frac{\underline{Y}^* \cdot \underline{Z}_i}{||\underline{Y}^*|| \; ||\underline{Z}_i||} .$$

Through substitution and simplification, this can be written as

$$\cos \theta_i = \frac{\underline{Y}^* \cdot D_i D_i^- \underline{Y}^*}{||\underline{Y}^*||} .$$

Recalling from Section II.B that $\underline{Y}^* \cdot \underline{Y}^* = 1$, and squaring both sides, $\cos^2\theta_i$ can be written as follows:

$$\cos^2\theta_i = \frac{||D_i \hat{\underline{\beta}}_i||^2}{||\underline{Y}^*||^2} = \frac{\underline{Y}^{*T} D_i^- D_i^T D_i D_i^- \underline{Y}^*}{1} .$$

Thus,

$$R^2_{\underline{Y}^*;D_i} = \cos^2\theta_i .$$

Matrix algebra then allows the following conversion for $R^2_{\underline{Y}^*;D_i}$:

$$D_i^T D_i D_i^- = [(D_i D_i^-)^T D_i]^T = [D_i D_i^- D_i]^T = D_i^T$$

$$R^2_{\underline{Y}^*;D_i} = \underline{Y}^{*T} D_i^{-T} D_i^T \underline{Y}^* = \underline{Y}^* (D_i D_i^-) \underline{Y}^*$$

Noting that $D_i D_i^-$ is symmetric, and letting

$$B_i = D_i D_i^- ,$$

we can write

$$R^2_{\underline{Y}^*;D_i} = \underline{Y}^{*T} B_i \underline{Y}^* .$$

Thus, the COF can be re-expressed as:

$$\min_{\underline{Y}^*} \max_i \underline{Y}^{*T} B_i \underline{Y}^* \qquad \text{(COFQ)}.$$

C. A GAME INTERPRETATION OF THE COFQ

The COFQ can be viewed as a game between a person and nature. The person tries to choose the B_i matrix that will maximize R^2, and nature plays the part of an antagonist who wants to create the least favorable $\underline{Y}$ to be predicted by that B_i.

The game is visualized as depicted in figure three. For a fixed $\underline{Y}$, the regression 'black box' (labled '1') determines the $\hat{\underline{\beta}}$ whose coefficients express $\underline{Y}$ as a linear combination of the 'best' subset of independent variables (columns of matrix X), D_i.

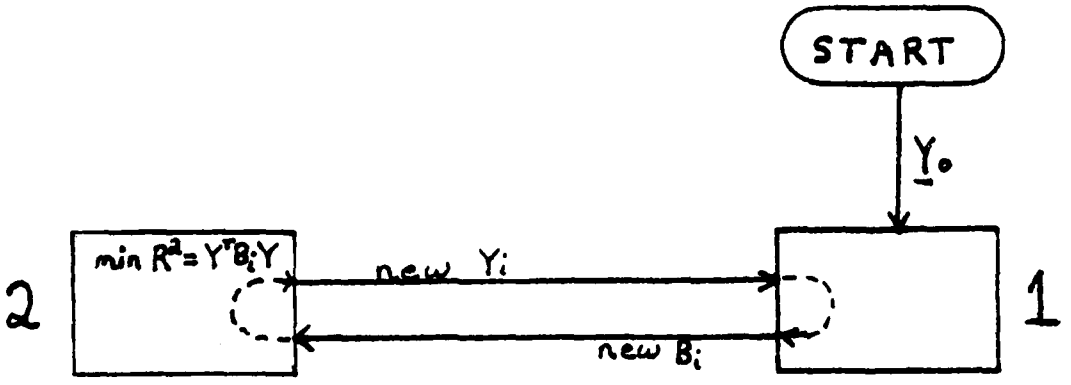


Figure 3. Game Interpretation of COFQ

Nature upon seeing the subspace represented by the matrix B_i, antagonistically chooses the worst $\underline{Y}$ in E^n so as to thwart the regression model's validity. This is represented in figure three as the box labled '2'.

At first glance, this view of the problem hints that there exists an iterative solution to the COFQ. Movement towards such a solution might be measured by the generation of a sequence of successively lower values of R^2 for the COFQ. However, obtaining an optimal $\underline{Y}$ is dependent on whether the sequence generated is convergent. Using Zangwill's general convergence theorem, it can be shown that this process will not generate a convergent sequence [Ref. 5]. For every B_i selected, nature will find a vector $\underline{Y}$ orthogonal to that corresponding subspace. It is the case that in a finite dimensional vector space, for any subspace with rank less than the vector space itself, there exists a vector orthogonal to that subspace. This is shown in figure 4 for 3-dimensions, where the subspace formed by x_1 and x_2 is of dimension 2. Vector $\underline{Y}$ is normal to the x_1, x_2, plane.

D. FORMULATION AS A NON-LINEAR PROGRAMMING PROBLEM (NLP) FOR OPTIMIZATION

In Section III.B, it was shown that the COF has an equivalent representation using a quadratic form for the objective function. The matrix B_i in the COFQ form is a square symmetrical matrix of size (n x n), and has rank k. In general, a quadratic function F(x) has the form

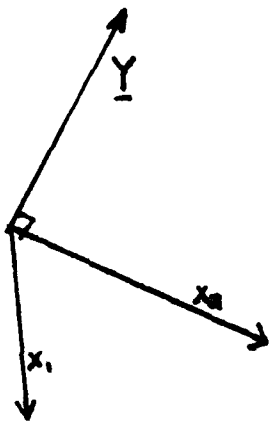


Figure 4. Y Orthogonal to Subspace

$$F(x) = \frac{1}{2}\underline{X}^T G\underline{X} + \underline{C}^T\underline{X} + \alpha$$

for a constant matrix G, column vector $\underline{C}$, and scalar α (multiplication by 1/2 is included in the quadratic term to avoid the appearance of a factor of two in the derivatives). The quantity 'G' is referred to as the Hessian matrix of F, which is the matrix of second partial derivatives. Thus, B_i can be characterized as the Hessian matrix of a quadratic form.

1. Modeling of the COFQ in Stages

The solution procedure for the COFQ can be broken down into several stages.

(a) For a particular B_i, optimize the quadratic form to find the minimum $\underline{Y}^*$.

(b) Use an algorithm for optimal subset selection to look at some regions only and thereby avoid optimizing over

all $\binom{p}{k}$ possible B_i. This is based upon the reasoning that some B_i (along with the associated constraints to be specified in the next section) will define convex regions which are larger than others and that the larger regions will produce a $\underline{Y}$ which is 'worse' than smaller regions will. Amongst those regions over which optimization was done, select the $\underline{Y}^*$ corresponding to the smallest R^2 value as the global minimum. These steps are addressed in the remainder of this chapter.

2. Modeling as a NLP

This section addresses stage (a); optimizing with a particular B_i to find the minimum $\underline{Y}^*$. Each optimal $\underline{Y}^*$ corresponding to a B_i is a local optimum for the overall COFQ problem.

Modeling this stage as a NLP (non-linear programming) problem, the objective function becomes:

$$\min \underline{Y}^{*T} B_i \underline{Y}^* \ .$$

From Section III.A, the constraints require $\underline{Y}$ to be of unit length and that $\underline{Y}$ be obtained through use of standardized regression coefficients. Assume the latter is met, hence the use of $\underline{Y}^*$ in lieu of $\underline{Y}$. The unit length requirement was stated as

$$\underline{Y}^{*T} \underline{Y}^* = 1 \ ,$$

and the latter constraint was represented by

$$\underline{1}^T\underline{Y} = 0 .$$

The NLP now becomes:

$$\min \underline{Y}^{*T} B_i \underline{Y}^*$$

$$\text{s.t. } \underline{Y}^{*T}\underline{Y}^* = 1$$

$$\underline{1}^T\underline{Y}^* = 0$$

$$y_i^* \text{ free; } i = 1,\ldots,n$$

Note that the y_i^* are free variables, otherwise the standardization requirement,

$$\sum_{i=1}^{n} y_i = 0 ,$$

could not be met.

The characterization of this NLP is that of a NLP with a non-linear equality constraint, referred to in the literature as a NEP [Ref. 6]. The constraint is a quadratic form for which the Hessian is the identity matrix. In searching for the locally optimal $\underline{Y}^*$, the optimization search must move along a quadratic surface at unit length from the origin of a n-dimensional hypersphere. (Recall from Section II.B that the intercept of the regression equation is zero.) This may be visualized in 3-dimensions as depicted in figure 5. This

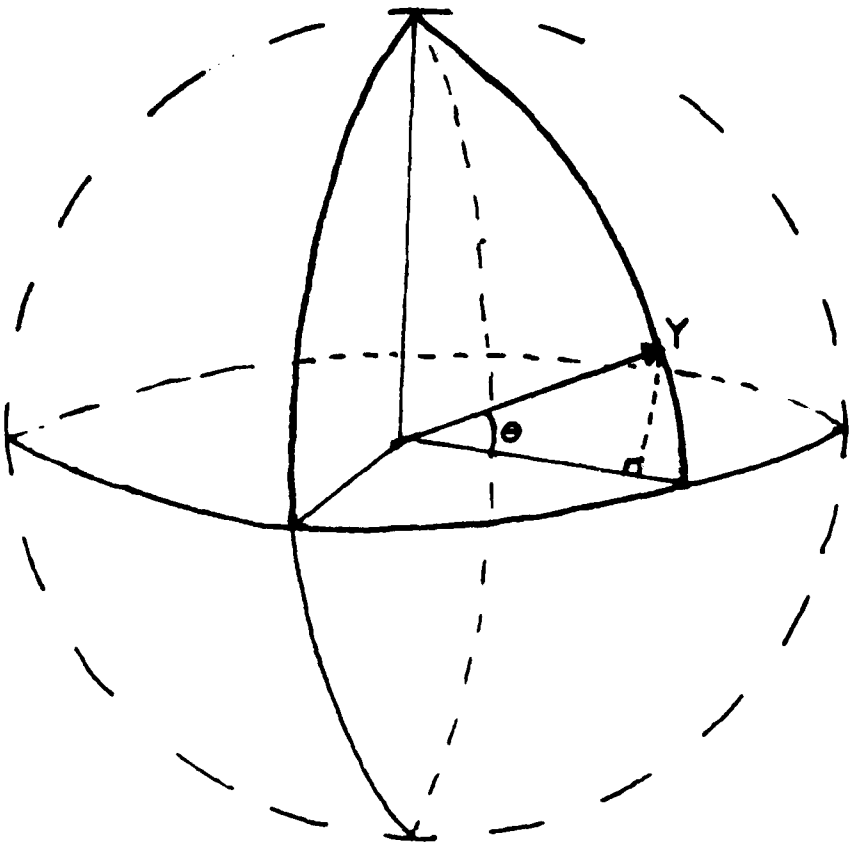


Figure 5. Movement of Y along Quadratic Surface

constraint requirement to move along a quadratic surface creates a non-convex NLP, which correspondingly increases the optimization complexity. Search direction methods for unconstrained nonlinear optimization are nonapplicable. Further, reduced-gradient and gradient-projection methods developed for problems with nonlinear constraints will probably fail due to the non-convexity.

The requirement to solve many NLPs as the B_i are changed necessitates an optimization procedure that is general in nature.

IV. PROBLEM FORMULATION TWO

The inherent difficulty with formulation one prompted a search for a computationally simpler model. By investigating the geometrical representation of the COF, could a linear model be built that, when solved, would solve the original problem? If an exact solution to the original problem was not reached, could an approximation be found that meets the 'practical need' for the COF value as expressed in the Introduction of this paper?

Such a linear model could utilize the extensively known results in linear programming; especially capitalizing on the speed of computation using existing linear programming algorithms.

A similar methodology as was used in Section III.D to break the COFQ problem into stages, will be used in the solution approach to the COF. One stage is optimizing to find the minimum $\underline{Y}^*$s locally through the use of a surrogate objective function. The second stage is either determining the global minimum $\underline{Y}^*$ through enumeration, or selecting the local $\underline{Y}^*$ most attractive as the answer (thereby approximately solving the COF).[4]

---

[4]Enumeration seems economically feasible into the neighborhood of approximately 14 candidate predictor variables.

A. GEOMETRICAL INTERPRETATION

Let $\underline{a}^T = (a_1,\ldots,a_n)$ be a non-zero vector. Consider the vectors $\underline{X}$ that satisfy

$$\underline{a}^T\underline{X} = d$$

for some scalar d. The set of $\underline{X}$ that satisfies this is defined to be a hyperplane. The vector $\underline{a}$ is termed the normal to the hyperplane, and the normalized vector

$$\frac{\underline{a}}{||\underline{a}||}$$

which has Euclidean length unity, is said to be the 'unit normal' to the hyperplane.

One can think of a hyperplane as a shift from the origin of the (n-1)-dimensional subspace orthogonal to $\underline{a}$ [Ref. 7]. This can be seen in figure 6. Note that if d = 0, the hyperplane (subspace) passes through the origin. This can be seen for three dimensions in figure 7.

Recall from Section I.B that the column vectors of the matrix X (without dummy variable) used in linear regression define a n-dimensional finite vector space called E^n. Further recall that combinatorial sets of k such vectors span subspaces of E^n. If the assumption of linearity is valid for the regression model, then interpretation of such subspaces as being linear is valid. Thus, let each normalized column

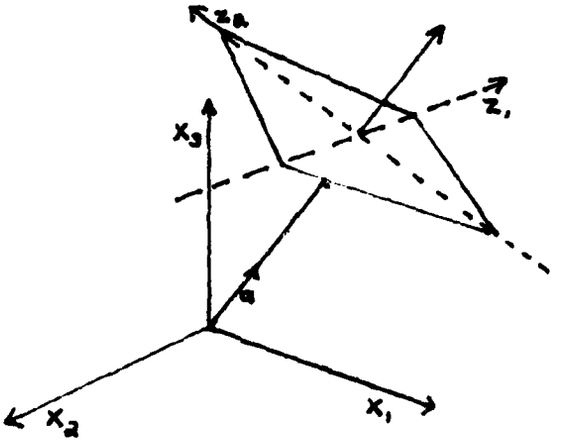


Figure 6. Hyperplane in 3-Dimensional Space

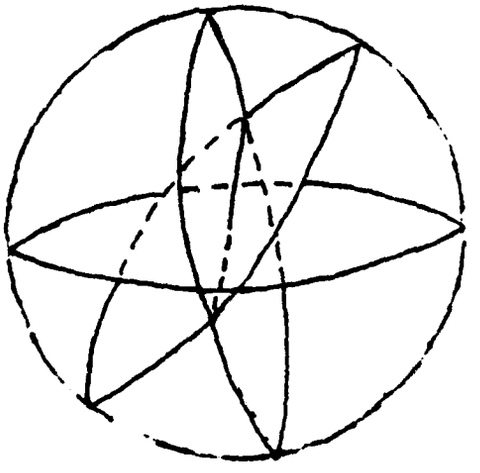


Figure 7. Hyperplanes through Center of a Sphere

vector of X represent the 'unit normal' to a (n-1)-dimensional hyperplane. Next, view each such hyperplane as a constraint of the form[5]

$$\underline{a}_i^T \underline{X} = d ; \quad i = 1,2,\ldots,p$$

where i denotes the i^{th} column of the data matrix of n-observations on p-predictor variables, and $\underline{X}^T = (x_1, x_2, \ldots, x_n)$.

Thus, the general equation for the i^{th} constraint can be written as follows:

$$a_{1j}x_1 + a_{2j}x_2 + \ldots + a_{nj}x_n = d$$

where

$$d = 0 \quad \text{for} \quad i = 1,\ldots,p .$$

Collectively, these p constraints (or hyperplanes) intersect within the hyperspace and form convex polytopes; specifically, cones. An elementary example of this is shown in figure 8. Any vector in E^n originating at the origin (to include $\underline{Y}$, the vector of interest) will geometrically lie within some such cone as defined by a set of hyperplanes.

---

[5]A notation switch has taken place to enable an easy transition to commonly used linear programming notation. This $\underline{X}$ does not represent the predictor variables, but rather real numbers to be determined.

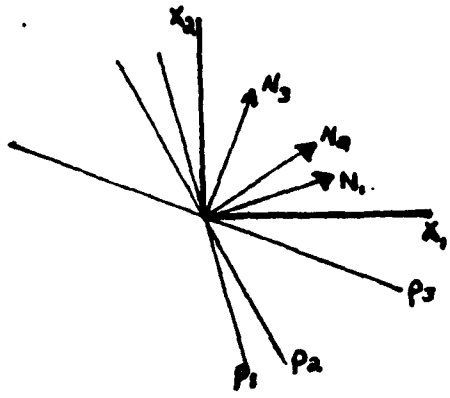


Figure 8. Convex Regions Formed in 2-Dimensions

B. ALGEBRAIC MODEL

Consistent with linear programming notation, the p-constraints defined by their unit normals can be written as follows:

$$
\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= 0 \\
a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= 0 \\
&\vdots \\
a_{p1}x_1 + a_{p2}x_2 + \ldots + a_{pn}x_n &= 0 \ .
\end{aligned}
$$

This system of constraints can be put into the standard form $A\underline{X} = \underline{b}$, where $\underline{b} = \underline{0}$. Matrix A is the matrix of

coefficients and is actually the transpose of the normalized column vectors from the independent (predictor) variable data matrix.

Using the concepts of halfspaces, specifying $\leq$ or $\geq$ in each relationship above determines on which side of each constraint a point in E^n lies. Putting the system of constraints into standard form will then require a slack (surplus) variable for each constraint.

C. CHARACTERIZATION OF THE CONES

Since the regression model projects the dependent variable $\underline{Y}$ onto the subspaces (hyperplanes), we want to know for any point (vector) $\underline{Y}$ which hyperplanes are the closest to $\underline{Y}$. This can be rephrased as asking, "Within which cone (convex polytope) does Y lie?"

Consider some arbitrary point (i.e., vector originating from origin) $\underline{X}_0^T = (x_1, x_2, \ldots, x_n)$ in the hypersphere. Further, assume that the points are constrained in the distance they can be from the origin, so as to not have an infinite ray.

Now, scanning out from $\underline{X}_0$, it can be seen that some hyperplanes are 'closer' to $\underline{X}_0$ than are others ('closest' defined by the smallest angle θ_i between $\underline{X}_0$ and any vector in a particular hyperplane). Define the set of hyperplanes which are closest to $\underline{X}_0$ when considering all directions as 'bounding hyperplanes', and the region bounded as the cone within which $\underline{X}_0$ lies; refer to such a cone as 'hole H_i'. The specific sizes and quantity of holes created by the intersecting

hyperplanes depends on the values given in the problem data. As we scan out from $\underline{X}_0$, we know for certain that the closest hyperplane is one of the walls. Note, however, that we cannot extend this argument to identify the other walls of the hole in which $\underline{X}_0$ lies. Thus, the question remains in identifying the walls of the hole (bounding hyperplanes). The concepts for reduction of linear inequalities presents an answer to this question [Ref. 7].

1. Reduction of Inequalities

Suppose for any arbitrary point $\underline{X}$ in E^n (subject to $\underline{X} \leq$ upper bound) all given hyperplanes are examined. Then there exist hyperplanes 'exterior' to the convex region defining H_i. These exterior hyperplanes, when viewed as constraints are non-binding. As an elementary example of this, a hole H_i is depicted in figure 9 corresponding to the following system of constraints.

$$\underline{a}_1^T\underline{X} \leq 0 \qquad \text{hyperplane } p_1$$

$$\underline{a}_2^T\underline{X} \geq 0 \qquad \text{hyperplane } p_2$$

$$\underline{a}_3^T\underline{X} \geq 0 \qquad \text{hyperplane } p_3$$

$$\text{L.B.} \leq \underline{X} \leq \text{U.B.}$$

As can be seen from figure 9, the third constraint is exterior to H_i (the hole identified by the letter i) and can

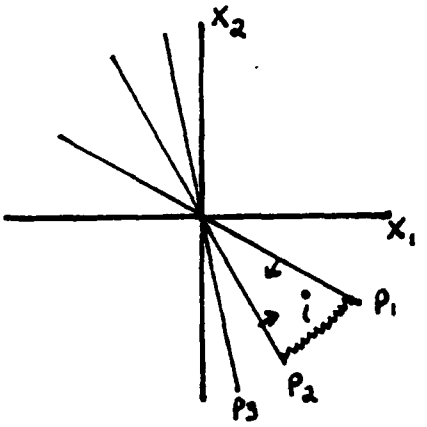


Figure 9. 2-Dimensional Depiction of Non-Binding p

be termed 'redundant'. This redundant constraint can be eliminated without changing the solution set.

Now, suppose a convex region is specified by adding slack variables, and putting the system of equations in standard form. Thus (assuming artificial variables have been forced out), we get:

$$\underline{a}_1^T\underline{X} + s_1 = 0$$

$$-\underline{a}_2^T\underline{X} + s_2 = 0$$

$$-\underline{a}_3^T\underline{X} + s_3 = 0$$

$$\underline{X} \geq |\text{L.B.}|$$

$$\underline{S} \geq 0 \ .$$

In standard form, this constraint redundancy is reflected in that s_3 is a 'non-extremal' variable and hence it, together with the third constraint, can be eliminated. This special example generalizes to higher dimensional problems. In general, redundant inequalities show up as having non-extremal slack variables.

Here the L.B. is any arbitrary positive real number sufficiently big to avoid roundoff problems, yet to prevent convergence at the origin. The following algorithm, using simplex techniques, will identify the binding constraints (walls) for any hole.

2. Algorithm: Removal of Redundant Constraints

Assume that some combination of 'less than or equal to' and 'greater than or equal to' inequality signs for the p constraints are specified. Call this system of p inequalities the 'resource constraints' and put it in standard form.

a. Find a feasible solution: min $\sum_{i=1}^{p} A_i$, where A_i is the artificial variable corresponding to the i^{th} constraint.[6]

b. Let $S = \{s_1, s_2, \ldots, s_p\}$; the set of slack variables.

c. Select s_i from S.

---

[6]Failure to find a feasible solution for that combination of ($\geq$,$\leq$) signs implies that a convex region is not defined, and therefore that combination can be ignored.

d. Min s_i, subject to resource constraints and variable bounds.

e. If s_i is non-extremal ($s_i > 0$), place index i in set R.

f. If set S has been exhaustively examined, go to Step g; otherwise increment i and go to Step c.

g. Remove or 'fix' all constraints i s.t. i ϵ {R}.

The result of this is that the boundaries of the convex region within which a specified point (vector) lies have been identified. Any optimization need only consider this subset of the resource constraints.

D. OPTIMIZATION FOR LOCAL MINIMA

The COF seeks a $\underline{Y}$ such that R^2 is minimized. For a given $\underline{Y}_0$ and the corresponding cone within which $\underline{Y}_0$ lies, this minimum is approached as $\underline{Y}$ moves away from the nearest hyperplane. Recall that as the angle of projection increases, R^2 decreases. However, past a certain position, the angle of projection between $\underline{Y}$ and a different hyperplane will decrease enough such that the regression will select this second hyperplane to project $\underline{Y}$ onto instead, since a higher R^2 value can be obtained.

If this process were to occur between $\underline{Y}$ and each of the walls of the hole simultaneously, an 'equilibrium point' for $\underline{Y}$ would be reached. This equilibrium point is defined to be the center of the hole. This can be visualized in 3-dimensions in figure 10.

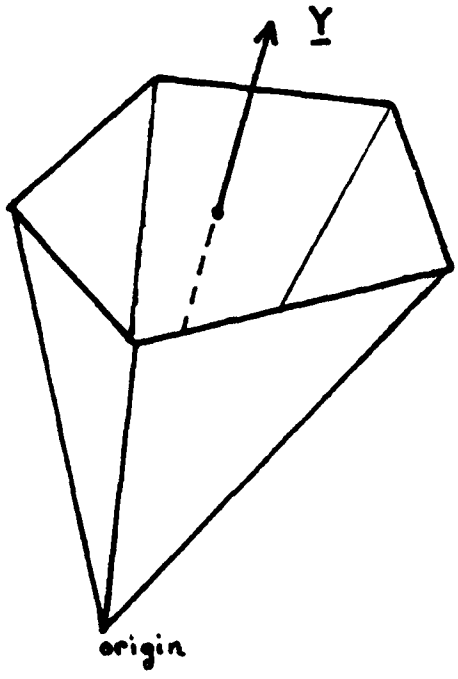


Figure 10. Vector Y at Equilibrium Point

It is at this equilibrium point where if $\underline{Y}$ were projected onto any of the walls, that the angles of projection would be the 'worst'; we are maximizing the minimum angle. The $\underline{Y}$ representing this point, and normalized to be unique, is the worst $\underline{Y}$ to be predicted using regression for this region of E^n.

The search for a local minimum $\underline{Y}$ thus becomes a search for a vector that originates at the origin and passes through the center of the hole. Considering the hole in the context of a convex region, we are searching not for a solution at an extreme point, but rather at the center of the region.

Viewing the previous figure from 'topside', we see $\underline{Y}$ as a point in the center of a (n-1)-dimensional convex region.

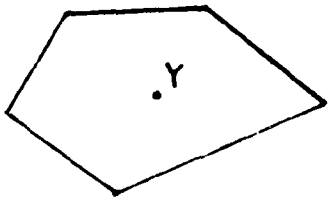


Figure 11. Top View of 2-Dimensional Convex Region

Such an equilibrium point can be approximated by optimizing using slack associated with the extremal constraints representing the hole. The objective function for optimization can be formulated as

$$\max_{\underline{X}} \left[\min_{i} \underline{C}_i^T \underline{X}\right] ,$$

where $\underline{C}_i$ is included for all $i \notin \{R\}$. Note each $\underline{C}_i$ contains only one non-zero coefficient--that coefficient representing the slack variable for the i^{th} resource constraint. The final $\underline{X}$ obtained is the optimal value of $\underline{Y}$.

The original form of the COF required $\underline{Y}^T\underline{Y} = 1$ (i.e., unit length), and that for the COFQ this requirement created a non-convex NLP. In figure 12 for 3-dimensions, it can be seen that there exists a hyperplane 'tangent' to this quadratic

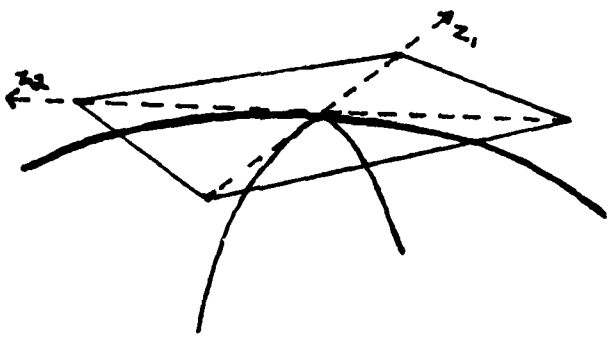


Figure 12. Hyperplane Touching Surface at One Point

surface such that the closest point on the hyperplane to the origin is the optimal $\underline{Y}^*$. Then the vector $\underline{Y}^*$ is the orthogonal unit normal to the hyperplane.

Now, let the initial value of $\hat{\underline{Y}}$ ($\hat{Y}$ is normalized) specify a bounding hyperplane to the hole. For each new value of $\underline{Y}$ a new bounding hyperplane can be specified, thereby linearly approximating the quadratic surface in successive increments as $\underline{Y}$ moves towards the center of the hole (optimal $\underline{Y}$).

Using matrix $\underline{A}$ with the understanding that it contains only the extremal resource constraints, the optimization problem can be formulated as

$$\max_{\underline{X}} \left[\min_{i} \underline{C}_i^T \underline{X}\right]$$

$$\text{s.t. } \underline{A}\underline{X} + \underline{I}\underline{S}_A = 0$$

$$\underline{a}^{kT}\underline{X} \leq 1$$

$$\underline{S}_A \geq 0$$

$$\underline{X} \text{ free}$$

where $\underline{a}^k$ = values of X for the $(k-1)^{th}$ iteration, and $\underline{Y}$ is represented by the current value of $\underline{X}$. For each hole, the initial $\hat{\underline{Y}}$ is the normalized value of the BFS to the L.P. which was reached using the algorithm in Section C.2. The values of $\hat{\underline{Y}}$ for the $(k-1)^{th}$ iteration become the coefficients of $\underline{a}^k$ (k^{th} iteration). Note that $\underline{Y}$ is normalized between each iteration. The quantity 'I' is an identity matrix and $\underline{S}_A$ is the column vector of slack variables corresponding to the extremal resource constraints (hence the use of subscript A).

This problem can be rewritten in final form as the L.P.:

$$\max \quad Z$$

$$\text{s.t.} \quad A\underline{X} + I\underline{S}_A = 0$$

$$Z \leq \underline{C}_i^T \underline{X} \qquad i \notin \{R\}$$

$$\underline{a}^{kT}\underline{X} \leq 1$$

$$\underline{S}_A \geq 0$$

$$\underline{X} \text{ free.}$$

E. OPTIMIZATION FOR A GLOBAL MINIMUM

Recall from Section IV.B that the direction of the inequality signs for the resource constraints determines which convex

set is the feasible region. We can see that there are 2^p such sets to consider for local optimization in the search for a global minimum to the COF.

When the number of candidate predictor variables is small, say up to about p = 14, enumeration is economically practical. For problems with p larger than this, only holes which look larger than most of the other holes could be optimized as 'candidates' for the global minimum. Such a selective optimization would require some type of global information prior to any optimization. A procedure to find such information was not found. This impasse led to the algorithm presented below. This algorithm, when given a local optimum, searches adjacent holes for a better minimum. If none better is found, it stops; otherwise, it uses the best hole and continues to search from there. In theory it might search all 2^p holes, but this is doubtful for real world problems (also, an iteration stop could be put in if desired). Note that the solution is still local in nature, although it may be the global solution. A heuristic approach may have to be developed to decide which hole to use to begin such a search.

1. Algorithm: Searching the Neighborhood about Y Min

An algorithm to search the neighborhood about a local minimum in an effort to find the global solution, and if not, to find a better minimum (larger Z value), is as follows:

a. Select some combination of inequality signs ($\leq$,$\geq$) for the resource constraints.

b. Get a BFS (basic feasible solution), remove redundant constraints, and then maximize Z (Z defined in previous section).

c. Record extremal constraints as set E (along with the direction of the inequalities), and record the value of $\underline{X}^*$ and Z.

d. For the k^{th} constraint of E, reverse the inequality (multiply the constraint by -1). Let J = {k^{th} constraint, reversed} $\cup$ {all resource constraints - k^{th}}. Optimize over J as in Step b. If $Z_k > Z$, record as in Step c (use label other than E), and set 'NEXT' = k.

e. If all elements of set E have been exhaustively examined, go to Step f; otherwise, increment k and go to Step d.

f. If no Z for set E is better than the initial Z, STOP. Otherwise, let J = {constraint 'NEXT'} $\cup$ {all resource constraints -'NEXT', reversed}. Fix 'NEXT' so as to not reverse its inequality sign again. Go to Step b.

This results in a minimum that although is local, can be considered the best solution possible in that region of the hypersphere.

Using the algorithms presented and the solution obtained, the analyst can compute the corresponding R^2 value by forcing the regression or using the equations in Section

III.B. Thus, he now has a relative measure of his model's applicability through comparison of the R^2 value pertaining to his prediction equation (obtained through least squares regression), to the lowest R^2 obtainable for that set of candidate predictor variables.

V. SUMMARY

The empirical model builder, in utilizing R^2 for a measure of 'goodness of fit', needs information concerning the quality of this statistic. This paper has addressed this problem utilizing optimization techniques to help the model builder assess the amount of confidence that can be placed in a R^2 value pertaining to a particular set of candidate predictor variables.

The linear programming algorithms presented offer a practical, fast, and cost effective methodology to search the hyperspace in which the linear regression model will operate. The lowest value of R^2 (globally) achievable for a particular set of cost data can be found when the number (p) of candidate predictor variables is small. Whereas, when the number of variables nears fourteen or more, a local minimum must be accepted due to computational costs inherent in the presented methodology.

LIST OF REFERENCES

1. Miller, B. M. and Sovereign, M. G., Parametric Cost Estimating with Applications to Sonar Technology, unpublished paper, Naval Postgraduate School, Monterey, California, 1973.

2. Clemson University Department of Mathematical Sciences Technical Report 350, On the Degree of Inflation of Measures of Fit Induced by Empirical Model Building, by T. B. Edwards and K. T. Wallenius, August, 1980.

3. Wallenius, K. T., Studying the Distribution for the Coefficient of Determination, unpublished research notes presented as Seminar topic at Naval Postgraduate School, Monterey, California, March, 1982.

4. Koerts, J. and Abrahamse, A. P. J., On the Theory and Application of the General Linear Model, Rotterdam University Press, 1971.

5. Zangwill, W. I., Nonlinear Programming, Prentice-Hall, 1969.

6. Gill, P. E., Murray, W., and Wright, M. H., Practical Optimization, Academic Press, 1981.

7. Luenburger, D. G., Introduction to Linear and Nonlinear Programming, Addison-Wesley, 1973.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Technical Information Center Cameron Station Alexandria, Virginia 22134	2
2.	Library, Code 0142 Naval Postgraduate School Monterey, California 93940	2
3.	Prof. G. T. Howard, Code 55Hk Operations Research Dept. Naval Postgraduate School Monterey, California 93940	2
4.	Prof. M. G. Sovereign, Code 74Zo C^3 Academic Group Naval Postgraduate School Monterey, California 93940	2
5.	Prof. K. T. Wallenius Dept. of Mathematics Clemson University Clemson, S. Carolina 29631	2
6.	CAPT Jeffery D. Vick, USMC Route 2, Box 201c Aitkin, Minnesota 56431	3
7.	Prof. G. G. Brown, Code 55Bw Operations Research Dept. Naval Postgraduate School Monterey, California 93940	1
8.	CAPT Dan Bausch, USMC Headquarters Marine Corps (I & L) Washington, D. C. 20380	1